Aihemallinnus hybridin mediatapahtuman ja merkitysten kierron tutkimuksessa

Toivanen, Pihla

2020-03

Toivanen , P , Huhtamäki , J , Valaskivi , K & Tikka , M 2020 , ' Aihemallinnus hybridin mediatapahtuman ja merkitysten kierron tutkimuksessa ' , Media & viestintä : kulttuurin ja yhteiskunnan tutkimuksen lehti , Vuosikerta. 43 , Nro 1 , Sivut 1-20 . https://doi.org/10.23983/mv.91078

http://hdl.handle.net/10138/314654
https://doi.org/10.23983/mv.91078

Artikkeli

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Aihemallinnus hybridin mediatapahtuman ja merkitysten kierron tutkimuksessa

Laskennallisten menetelmien käyttö viestintätieteissä on viime vuosina herättänyt kasvavaa kiinnostusta. Menetelmien soveltaminen vaatii kuitenkin sekä osaamista että materiaalisia resursseja, kuten palvelimia ohjelmien suorittamiseen. Tarkastelemme tässä artikkelissa Twitter-aineiston laskennallista keruuta sekä erityisesti sen analyysiä erään ohjaamattoman koneoppimismenetelmän, aihemallinnuksen avulla. Keräsimme artikkelin empiirisen aineiston maaliskuussa 2019 twiiteistä, jotka liittyivät Uuden-Seelannin Christchurchissa moskeijoihin suunnattuihin terrori-iskuihin. Tarkastelemme tässä artikkelissa, mitä ohjaamaton koneoppiminen voi antaa hybridin mediatapahtuman monimenetelmälliselle tutkimukselle, jossa mediaetnografinen tutkimusote on lähtökohta. Toteutimme empiiriselle aineistollemme aihemallinnuksen sen yleisten vaiheiden eli aineiston esikäsittelyn, mallinnuksen sekä tulosten tulkinnan mukaisesti. Aihemallinnuksen avulla tunnistimme 15 Christchurch -aineistosta nousevaa aihetta 148 816 twiitistä. Aihemallinnus tarjosi tavan tiivistää suurta aineistoa ja lähtökohtia tutkimuksen seuraavalle, laadulliselle vaiheelle. Käsittelemme artikkelissa aihemallinnuksen lisäksi myös laskennallisten menetelmien käyttöön sekä aineistonkeruuseen liittyviä rajoitteita.

AVAINSANAT: laskennalliset menetelmät, aihemallinnus, hybridi mediatapahtuma

Mediatapahtumien tutkimus ja teoria ovat viime vuosina pyrkineet ymmärtämään erityisesti disruptiivisten eli hajottavien tapahtumien erityisluonnetta hybridissä mediaympäristössä (vrt. Chadwick 2013; Valaskivi ym. 2019), käyttäen hybridin mediatapahtuman käsitettä (Vaccari ym. 2015; Sumiala ym. 2016; Sumiala ym. 2018, 15). Hybridissä mediaympäristössä sekoittuvat tuotannon ja vastaanoton käytännöt tavalla, jossa sekä ammattimaisesti vakiintuneiden mediaorganisaatioiden käytännöt että sosiaalisen median käytännöt muokkaavat toisiaan ja

syntyviä merkityksiä (Chadwick 2013). Hybridissä ympäristössä korostuvat inhimillisten ja ei-inhimillisten toimijoiden vuorovaikutus (Latour 1993; Sumiala ym. 2018, 17) sekä laaja, ylirajainen viestinnän mahdollisuus ja merkitysten tuotanto (Kraidy 2005; Sumiala ym. 2018, 2). Media- ja viestinnäntutkimuksen kentällä yleinen ja yhteinen ymmärrys lienee, että muuttuneen mediaympäristön dynamiikkojen ja aineistojen ja merkitysten kierron (Valaskivi & Sumiala 2014) ymmärtäminen edellyttää monimenetelmällistä otetta, joka yhdistää erilaisia lähestymistapoja (Laaksonen ym. 2017; Sumiala ym. 2016).

Yksi tällainen monitieteinen yritys on Hybrid Terrorizing -tutkimuskonsortio (2018–2021), joka Suomen Akatemian rahoittamana tarkastelee terroristista väkivaltaa suhteessa mediatapahtuman teoriaan (esim. Dayan & Katz 1992, 1–24; Dayan 2010; Eide ym. 2008; Katz & Liebes 2007; Liebes 1998; Nossek 2008; Rothenbuhler 2010). Hankkeen tavoitteena on muun muassa kehittää mediatapahtuman tutkimukseen soveltuvaa, laadullisia ja määrällisiä menetelmiä yhdistävää lähestymistapaa. Hankkeessa ollaan kiinnostuneita muun muassa siitä, miten hajottavia mediatapahtumia (Katz & Liebes 2007) voidaan ymmärtää paremmin laskennallisten menetelmien (Sumiala ym. 2016; Laaksonen ym. 2017) avulla vaikkakin mediaetnografisella otteella. Tällöin tavoitteena on, että laskennalliset menetelmät antavat suuntaa laadulliselle analyysille.

Tässä artikkelissa tarkastelemme hajottavan mediatapahtuman laskennallisen aineistonkeruun metodologiaa, tapauksenamme Uudessa-Seelannissa Christchurchissa maaliskuussa 2019 kahteen moskeijaan suunnattu terrori-isku. Keskitymme erityisesti aineiston keruun ja ensivaiheen analyysin toteutukseen laskennallisia menetelmiä käyttäen ja pohdimme millaisia lähtökohtia nämä menetelmät tarjoavat laadulliselle tulkinnalle. Kun laskennallisen aineiston koko liikkuu miljoonissa dokumenteissa - tässä tapauksessa twiiteissä - on siitä hankala saavuttaa kokonaiskuva laadullisesti lähilukemalla. Sovellamme tässä artikkelissa aihemallinnusta (Blei ym. 2003; Ylä-Anttila 2018a; Bonilla & Grimmer 2013) ja erityisesti aineiston tiivistämistä aihemallinnuksen avulla (esim. Ylisiurua 2017). Aineiston tiivistämisen avulla valitsemme lähilukuun twiittejä tarkoituksenamme muodostaa jäsennelty kokonaiskuva hybridistä mediatapahtumasta. Kysymme millaisia mahdollisuuksia ja rajoitteita aihemallinnus tarjoaa hybridin mediatapahtuman tutkimuksen näkökulmasta, kun pyritään säilyttämään mediaetnografinen ote.

Tämä artikkeli koostuu kolmesta osiosta, joista ensimmäisessä käymme lyhyesti läpi mediatapahtuman teoriaa, kuvaamme Christchurchin terrori-iskuun liittyvät tapahtumat ja kerromme Twitter-aineiston laskennallisesta keruusta mediatapahtuman ollessa käynnissä. Seuraavassa osiossa pohdimme aineiston koon analyysille aiheuttamia haasteita. Tarjoamme ratkaisuksi yhtä ohjaamattoman koneoppimisen menetelmää, aihemallinnusta, ja esittelemme aihemallinnuksen periaatteita, aineiston analyysiprosessia ja tuloksia, sekä pohdimme aihemallinnuksessa tehtävien valintojen seurauksia tutkimukselle. Viimeisessä osiossa keskustelemme aihemallinnuksen roolista hybridin mediatapahtuman tutkimuksessa.

Disruptiivinen hybridi mediatapahtuma

Kun Daniel Dayan ja Elihu Katz julkaisivat kirjansa *Media Event. The Live Broadcasting of History* (1992), heidän keskeinen ajatuksensa oli analysoida television roolia seremoniallisissa, kansakuntaa yhdistävissä tapahtumissa. Teos tarkastelee median rituaalista luonnetta ja sen merkitystä sosiaalisen koheesion tuottamisessa tapahtumissa, jotka on ennalta suunniteltu niin tapahtumajärjestäjien kuin mediankin taholta. Dayan ja Katz jaottelevat seremonialliset tapahtumat kruunajaisiin, kilpailuihin ja voiton hetkien juhlimiseen. Dayanin ja Katzin alkuperäisessä teoriassa tällaiset etukäteen suunnitellut, seremonialliset tapahtumat oli nimenomaan erotettu yllättävistä uutistapahtumista.

Dayania ja Katzia onkin kritisoitu siitä, että he eivät näyttäneet näkevän uutistuotannon rituaalista luonnetta, ja siitä, että he keskittyivät vain televisioon, eivätkä tarkastelleet sitä monimediaista kokonaisuutta, joka mediatapahtumaan aina liittyy (esim. Sonnevend 2016). Elihu Katz päivitti yhdessä Tamar Liebesin kanssa mediatapahtuman teoriaa artikkelilla *'No More Peace!' How Disaster, Terror and War Have Upstaged Media Events* (2007), jossa he tarkastelivat terrorismia syyskuun yhdennentoista päivän iskujen jälkitilassa. Artikkelissaan Katz ja Liebes esittelevät disruptiivisen eli hajottavan mediatapahtuman käsitteen, jolla he viittaavat sellaisiin katastrofeihin, jotka häiritsevät yhteiskunnan toimintaa, tuottavat kriisin tunteen ja vaativat monenlaista kollektiivista, medioitunuttakin työstämistä normaalitilanteen palauttamiseksi. Terrorismi on toki ennalta suunniteltua, mutta sen suunnittelusta ovat tietoisia vain iskun tekijät.

Hybridin mediatapahtuman käsite pyrkii kehittämään Dayanin ja Katzin alkuperäistä teoriaa monimediaisuuden ja -kanavaisuuden osalta. Monimediainen ympäristö syntyi jo ennen internetiä, Yhdysvalloissa televisio- ja radiokentän laajentumisena 1960-luvulta lähtien ja muualla 1980-luvulta alkaen, kun kaapeli- ja satelliittitelevisio ja paikallisradiot monimutkaistivat ja lisäsivät tarjontaa huomattavasti. Kuitenkin juuri internetin ja erityisesti sosiaalisen median alustojen suosion kasvu muutti tuotannon ja vastaanoton suhteita, kasvatti teknologian merkitystä mediatapahtumien kollektiivisen luonteen rakentumisessa ja vahvisti mediaympäristön ylirajaisuutta. Samalla vanhempien, kansallisten mediainstituutioiden rooli ja mahdollisuus hallita mediatilaa ovat kaventuneet.

Dayanin ja Katzin alkuperäistä tutkimusta on kritisoitu myös siitä, ettei heidän tutkimuksensa kiinnity mihinkään empiiriseen aineistoon (esim. Sonnevend 2016). Sittemmin mediatapahtuman teoriaa on kyllä käytetty runsaasti erilaisten konkreettisten tutkimustapausten tarkasteluun (esim. Couldry ym. 2010; Kyriakidou 2008). Hajottavan mediatapahtuman hahmottaminen hybridissä mediaympäristössä on erityisen haastava tehtävä, jossa on tärkeää olla tietoinen siitä, että aineistojen tuottamisen tavat ja menetelmien valinnat vaikuttavat siihen, kuinka ymmärrämme tapahtuman merkityksen ja ulottuvuudet. Hajottavien mediatapahtumien tutkimuksen haasteellisuus liittyy muun muassa siihen, että ne ovat suunnittelemattomia uutismedian, julkisten toimijoiden, suuren yleisön – ja tutkijoiden – näkökulmasta. Mediaympäristön

luonteen vuoksi aineiston keruu täytyy aloittaa mahdollisimman varhain tapahtumien alettua. Jälkikäteen koottavissa oleva aineisto on luonteeltaan hyvin erilaista suhteessa siihen vähitellen hahmottuvaan uutisten, viestien ja merkitysten pyörteeseen, joka tapahtumien edetessä kiertää.

Mediatapahtumien tutkimus on pitkään painottunut laadullisiin menetelmiin (esim. Couldry ym. 2010). Nykyisessä mediaympäristössä toimijoiden, viestien ja representaatioiden moninaisuus ja suuri määrä kuitenkin kutsuvat laadullisten menetelmien tueksi laskennallisia välineitä. Olemme kehittäneet erilaisia monimenetelmällisiä tapoja hajottavan mediatapahtuman tutkimiseen hybridissä mediaympäristössä jo useamman vuoden ajan ja käsitteellistäneet tätä työtä hybridin mediatapahtuman käsitteen avulla (Sumiala ym. 2016; Sumiala ym. 2018). Tässä artikkelissa jatkamme tätä työtä ja pohdimme sitä, miten monimenetelmällisellä lähestymistavalla voidaan muodostaa ja analysoida laajaa ja mutkikasta aineistoa.

Tapaus Christchurch

Perjantaina 15. maaliskuuta 2019 Uuden-Seelannin Christchurchissa 28-vuotias mies hyökkäsi kahteen moskeijaan (BBC, 15.3.2019) ja surmasi yhteensä 51 ihmistä (CNA, 2.5.2019). Hyökkääjää pidetään äärioikeistolaisena ja hän julkaisi ennen hyökkäyksiä rasistisen dokumentin, jossa ilmoitti tulevan hyökkäyksensä kohteen.

Iskulla oli kaksi tälle tutkimukselle tärkeää erityispiirrettä: Ensiksi mediateknologian laaja käyttö, ja toiseksi iskun jälkipuinnin aikana levinnyt *nameless*-ilmiö. Mediateknologian rooli tapahtumassa oli kauhistuttavimmillaan, kun tekijä kuvasi ja välitti iskunsa reaaliaikaisesti Facebookiin. Sittemmin Facebook poisti tekijän tilin sekä ilmoitti yrittävänsä poistaa tallenteen kopiot (BBC, 15.3.2019), joskin video oli jo kopioitu lukuisiin muihin paikkoihin ja löytyy edelleen verkosta helposti. Namelessilmiö liittyy Uuden-Seelannin pääministerin Jacinda Ardernin toimintaan iskun jälkeen. Hän ilmoitti heti iskun jälkeen, ettei aio lausua ja siten toistaa tekijän nimeä, ja kehotti muitakin kiinnittämään huomionsa uhreihin ennemmin kuin tekijään. (Wahlquist 2019). Tähän julistukseen liittyy haaste tutkimukselle: Kuinka toteuttaa tutkimus niin, ettemme tule antaneeksi tekijälle lisää huomiota. Christchurchin hyökkääjä – muiden terroristien lailla – tavoitteli kuuluisuutta ja mahdollisimman laajaa näkyvyyttä asialleen. Tätä varten hän kirjoitti myös merkillisen manifestin, jonka monitulkintaiset kulttuuriset viittaukset suorastaan kutsuvat jatkuvaan analyysiin ja spekulointiin. Jotta emme antaisi hyökkääjälle tarpeetonta huomiota, aiomme tässä artikkelissa sivuuttaa kyseisen manifestin ja keskittyä analysoimaan tapahtuman muita ulottuvuuksia.

Kiskot rakentuvat liikkuvan ”datankäsittelyjunan” alle

Kun jotain yllättävää tapahtuu, alkaa aiheen käsittely sosiaalisessa mediassa välittömästi. Keskitymme tässä tutkimuksessa Twitteriin, jonka rooli erilaisten uutistapah-

tumien alkuvaiheessa on vähitellen muodostunut keskeiseksi sekä uutismedian että kansalaisten ja viranomaisten näkökulmasta (Bruno 2011; Vis 2013). Twitter on otollinen väline datan keräämiseen, koska yritys on kehittänyt kolmansille osapuolille teknisiä keinoja datan keräämiseen. Twitterin teknisten ominaisuuksien vuoksi sen dataan perustuvaa laskennallista tutkimusta tehdään paljon (vrt. Fiesler & Proferes 2018), itse asiassa huomattavasti enemmän kuin Twitterin käyttäjämäärät antaisivat aihetta. Tutkimusta puoltaa kuitenkin se, että juuri Twitterillä on sosiaalisen median alustoista erityinen rooli uutistapahtumissa. Twitter on siis otollinen ympäristö tarkastella terroriuutisia hybridin mediatapahtuman näkökulmasta.

Twitter tarjoaa ohjelmointirajapinnan (Application programming interface, API), jonka avulla dataa on mahdollista kerätä kahdella tavalla.[1] Dataa voi kerätä joko hakuehdot määrittelevillä pyynnöillä tai keräämällä datavirtaa reaaliaikaisesti. Pyyntöihin perustuvalla keräyksellä on mahdollista kerätä vain rajallinen määrä viimeaikaista dataa, minkä vuoksi Christchurchin iskun tapauksessa datavirran reaaliaikainen tallentaminen oli halutun aineiston suuren määrän takia ainoa tarkoituksenmukainen keräyskeino. Kuvaamme seuraavaksi käyttämämme aineiston keräystavan.

Aineiston kerääminen on valmisteltava huolella, jotta keräys saadaan käyntiin mahdollisimman nopeasti tutkittavan väkivaltaisen tapahtuman alettua. Kehitimme ja testasimme datakeräimen ennen mahdollista iskua siten, että voisimme käynnistää keräimen iskun tapahtuessa. Kun saimme tiedon Christchurchin iskusta ja päätimme ryhtyä keräämään dataa, määrittelimme ensimmäiset hakuehdot ja käynnistimme keräimen erityisesti tähän tarkoitukseen perustetulle palvelimelle. Keräin käynnistettiin noin puoli tuntia sen jälkeen, kun keräämisestä vastaava tutkija sai tiedon iskusta. Aikaeron vuoksi iskusta oli tässä vaiheessa kulunut jo useita tunteja. Datan keräämisestä vastaava tutkija päivitti hakuehtoja digitaalista mediaetnografiaa tekevien tutkijoiden ehdotusten mukaisesti ja ohjelman toimintaa seurattiin aktiivisesti tapahtuman akuutin vaiheen eli kahden viikon ajan. Hakuehtojen kehittely suoritettiin osana Facebookin pikaviestimessä käytyä keskustelua.

Lopulliseksi hakuehtolistaksi muodostui 'christchurch', 'christchurchmosqueattack', 'christchurchshooting', 'christchurchattack', 'JeSuisChristchurch', 'PrayForChristchurch', 'Al Noor', 'mosque shooting', 'mosque massacre', 'BrentonTarrant', 'NewZealandShooting', 'JeSuisHuman', 'HelloBrother', 'NewZealandTerroristAttack', 'NewZealandStrong'.

Kokemuksemme mukaan hakuehtojen määritteleminen on tasapainottelua aineiston edustavuuden ja määrän sekä tapahtuman tiheän kuvauksen välillä (Procter ym. 2013). Jos hakuehdot ovat liian yksinkertaiset, jää suuri osa aineistosta keräämättä. Ensimmäinen valitsemamme hakuehto oli 'christchurch'. Aihetunnisteen (hashtag) sijaan valitsimme pelkän sanan, joka osuu myös twiittien teksteihin ja jopa verkko-osoitteisiin, joissa sana esiintyy. Aiemmat tutkimukset ovat osoittaneet, että hajottavissa mediatapahtumissa Twitterin käyttäjät eivät aina käytä aihetunnistetta, vaikka twiittaavatkin aiheeseen liittyen (Tufekci 2014). Tämän vuoksi aineistonkeruu vain aihetunnisteen kautta vääristäisi väistämättä aineiston laajuutta. Akuutissa vaiheessa valtaosa twiiteistä, joissa 'christchurch' esiintyy, käsitteli terrori-iskua, mutta sekaan valikoitui myös muuta aineistoa. Hyvin yleiset hakuehdot tuottavat kattavan aineiston,

mutta epäolennaisen aineiston osuus kasvaa merkittävästi, mikä lisää työtä aineiston siivousvaiheessa.

Twitterin toimintaperiaate on yksinkertainen, mutta yksittäisistä twiiteistä kerättävissä oleva data on yksityiskohtaista. Listauksessa 1 on ote Jacinda Ardernin 15. maaliskuuta lähettämästä twiitistä Twitterin rajapinnan käyttämässä muodossa. Twiitin tärkeimpiä tietoja ovat teksti kokonaisuudessaan sekä uudelleenlähetysten ja suosikkimerkintöjen määrä. Twiitin lähettäjästä on saatavilla käyttäjän itsensä määrittelemät nimi, sijainti ja lyhyt kuvaus sekä seuraajien, seurattavien ja lähetettyjen twiittien määrä. Datasta löytyvät myös mahdolliset twiittiin lisätyt maininnat muista käyttäjistä, aihetunnisteet, linkit ja media-aineistot sekä tieto siitä, onko kyseessä uudelleenlähetys (retweet) tai vastaus (reply) twiittiin.

Listaus 1. Ote twiitin metatiedoista JSON-muodossa.

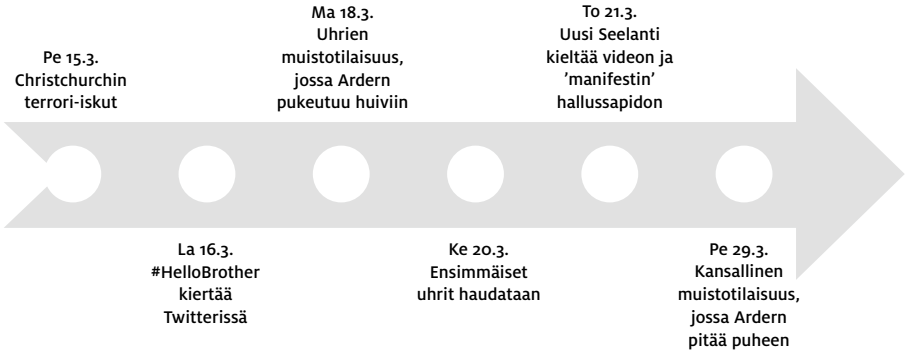
```
{
 ”created_at”: ”Fri Mar 15 03:33:02 +0000 2019”,
 ”id”: 1106397870628847617,
 ”full_text”: ”What has happened in Christchurch is an extraordinary
 act of unprecedented violence. It has no place in New Zealand.
 Many of those affected will be members of our migrant
 communities – New Zealand is their home – they are us.”,
 ”retweet_count”: 16453,
 ”favorite_count”: 44761,
 ”lang”: ”en”,
 ”user”: {
 ”id”: 22959763,
 ”name”: ”Jacinda Ardern”,
 ”screen_name”: ”jacindaardern”,
 ”location”: ”Auckland, New Zealand”,
 ”description”: ”Prime Minister of NZ. Leader @nzlabour. Won’t
 tweet what I ate for breakfast-make no promises beyond that.
 Auth by Rt Hon Jacinda Ardern, Parlt Buildings, WLG.”,
 ”followers_count”: 216900,
 ”friends_count”: 4408,
 ”statuses_count”: 6930,
}
```

Uudelleenlähetetyillä twiiteillä on merkittävä rooli reaaliaikaisessa datankeräyksessä. Jokainen uudelleentwiittaus nimittäin sisältää myös alkuperäisen twiitin yksityiskohtaiset tiedot. Tallentamalla uudelleenlähetyksiä on mahdollista kerätä varsin kattava aineisto suosituimmista alkuperäisistä twiiteistä, vaikka keräys aloitettaisiin viiveellä tapahtuman jälkeen. Tarkat tiedot twiittien levittäjistä ja leviämisen aikajanasta eivät kuitenkaan tallennu. Uudelleenlähetys on myös itsessään twiitti, mutta ei aina sisällä tekstiä.

Kumpi oli ensin, analyysi vai kokonaiskuva?

Datan keruu ja analyysi laskennallisten menetelmien avulla kietoutuvat tiiviisti toisiinsa. Kuten yllä toteamme, päivitimme datan keruun hakuehtoja tapahtuman kestäessä, mikä vaikutti lopullisen aineiston muotoon. Aineistonkeruun päätyttyä ryhdyimme hahmottamaan kokonaiskuvaa lähes kymmenestä miljoonasta twiitistä koostuvasta aineistosta. Lähdimme liikkeelle samanaikaisesti sekä aineiston ilmeisimmistä piirteistä että Christchurchin iskujen kiinnekohdista rakentaaksemme mediatapahtuman aikajanan (kuva 1). Hahmotimme laskennallisen analyysin keinoin aineiston koon ajallista kehitystä ja paikansimme piikit aineistokäyrällä empiirisiin tapahtumiin, kuten hautajaisiin ja muistotilaisuuksiin. Tämän perusteella rajasimme käsiteltävän aineiston kestoksi 15 päivää. Oma kokemuksemme sekä muu aiempi tutkimus osoittavat myös, että uutistapahtuma saa huomiota osakseen maksimissaan noin kaksi viikkoa (esim. Sumiala ym. 2018). Keräsimme valitsemamme viidentoista päivän ajalta 215 691 alkuperäistwiittiä, joiden uudelleentwiittauksia (retweet) oli kaikkiaan 8 204 495.

Kuva 1: Christchurchin väkivaltaisen mediatapahtuman aikajana.



Seuraavaksi ryhdyimme pohtimaan aineiston sisältöjen analyysia. Halusimme kartoittaa, mitkä aiheet ja toimijat nousevat aineistossa esille. Tämän vaiheen haasteina olivat sekä aineiston suuri koko että sen suhteellinen tuntemattomuus. Olimme innostavan kysymyksen äärellä: Miten tehdä laskennallista analyysia suurelle aineistolle, jonka sisältö on hakuehtoja ja kokoa lukuun ottamatta tuntematon tai tunnettu vain osin samaan aikaan toisaalla hankkeessa tehtyjen etnografisten havaintojen perusteella? Yhtäältä, jotta laskennallinen analyysi voidaan suunnitella ja toteuttaa, se vaatii eksakteja ohjeita. Eksploratiivinen analyysi ilman tutkimuskysymyksiä on myös mahdollista, jos vain on eksaktisti tiedossa, mihin huomio halutaan kohdentaa. Toisaalta tarkentavia ohjeita on vaikea esittää, koska ei vielä tiedetä, mitä tutkimuskysymysten näkökulmasta olennaisia ominaisuuksia aineisto sisältää. Vaiheen pulmallisuuden voi tiivistää toteamalla, että samaan aikaan kun aineiston laadullinen läpiluku on mahdotonta sen suuren koon vuoksi, laskennallisen analyysin aloittaminen ilman laadullista

harkintaa voi tuottaa vinoutuneen ymmärryksen aineistosta. Päätimme ryhtyä kartoittamaan aineiston sisältöjä koneoppimisen menetelmien avulla.

Koneoppiminen alkuvaiheen analyysin tueksi

Koneoppimista voidaan yksinkertaistetusti kuvailla kokemusten muuntamiseksi tiedoksi tai osaamiseksi (Shalev-Shwartz & Ben-David 2014, 19). Kokemukset tarkoittavat koneoppimismallille jonkinlaista *syötettä* eli opetusdataa, joka tässä tapauksessa koostuu twiiteistä. Lopputuloksena on jonkinlaista *osaamista*, jota yleensä käytetään koneoppimismallin rakentamisen jälkeen erilaisten tehtävien suorittamisessa.

Koneoppiminen sopii juuri sellaisiin tilanteisiin, joissa aineistoa on liikaa ihmisen analysoitavaksi tai kun ratkaistava tehtävä vaatii sopeutuvuutta erilaisiin syötteisiin (Shalev-Shwartz & Ben-David 2014, 21–22). Koneoppimisen lähtökohdat sopivat hyvin datan analyysiin kokonaiskuvan muodostamiseksi alkuvaiheessa. Koneoppiminen tässä tapauksessa tukee kokonaiskuvan muodostamista ja aineiston rajaamista esimerkiksi ajallisesti ja näin keskittymistä joihinkin mediatapahtuman osiin. Laskennallisen analyysin toistettavuus (*reproducibility*) helpottaa aineiston rajaamista ajallisesti, sillä saman analyysin voi helposti suorittaa useille eri aikajaksoille. Toistettavuus tukee myös analyysiprosessin kehittämistä vaiheittain. Koneoppimista ja tässä käyttämäämme aihemallinnusta on aiemminkin käytetty suuren keskusteluaineiston tiivistämiseen (Ylisiurua 2017).

Koneoppimisen käyttö viestinnän tutkimuksessa sijoittuu laajemmin laskennallisen yhteiskuntatieteen keskusteluun (Lazer ym. 2009; Cioffi-Revilla 2010), jossa on käsitelty esimerkiksi laskennallisten menetelmien objektiivisuusväitteitä, datan laatua koskevia kysymyksiä ja laskennallisten menetelmien etiikkaa (boyd & Crawford 2012). Objektiivisuusväitteiden mukaan humanistiset tieteet voivat suuria datamääriä käyttämällä asemoitua määrällisempiä sekä objektiivisempia metodeja käyttäviksi (boyd & Crawford 2012). boydin ja Crawfordin (2012) mukaan on kuitenkin virheellistä ajatella, että määrällisten tutkijoiden työ lähtökohtaisesti tuottaa faktoja ja laadullisten tulkintoja. Myös laskennallinen datankäsittely sisältää subjektiivisia valintoja, esimerkiksi ennen analyysiä tapahtuvassa esiprosessoinnissa (boyd & Crawford 2012; Huhtamäki ym. 2015). Aihemallinnuksen käyttäminen lähilukuun otettavien dokumenttien valinnassa on yritys ottaa huomioon laskennallisten menetelmien subjektiivisuus ja jättää näin tilaa tulkinnalliselle analyysille.

Ohjaamattoman koneoppimisen perusperiaatteet

Koneoppimismenetelmät voidaan jakaa ohjaamattomaan ja ohjattuun koneoppimiseen. Tekstianalyysin tapauksessa nämä molemmat perustuvat siihen, että aineisto koostuu dokumenteista. Ohjattu oppiminen perustuu opetusdataan, jossa osa aineistosta on esimerkiksi ihmisen luokittelemaa, ja koneoppimisen tavoitteena on toistaa

luokittelu koko aineistolle. Ohjaamattomassa oppimisessa erillistä luokiteltua opetusdataa ei tarvita, sillä ohjauksen sijaan ohjelma tunnistaa aineistossa piileviä säännönmukaisuuksia monimutkaisia laskutoimituksia tekemällä.

Aineistossamme ensivaiheen laskennallisen analyysin dokumentteja ovat alkuperäistwiitit. Dokumentit esitetään koneoppimisessa *piirrejoukkojen* avulla siten, että piirrejoukko on sama jokaiselle aineiston dokumentille, mutta piirteiden arvot vaihtelevat (Kotsiantis ym. 2006). Eräs yleinen piirrejoukko tekstianalyysissä on sanojen esiintyvyysmäärät (esim. Zhang ym. 2010). Jokaista dokumenttia vastaa siis lista, jossa on paikka jokaiselle sanalle, joka esiintyy vähintään yhden kerran koko aineistossa.

Seuraavassa esimerkissä havainnollistetaan twiittien muuntamista sanaesiintyvyyspiirteiksi. Taulukossa 1 esitetään aineistomme eräiden kahden twiitin alkuosat, taulukossa 2 niistä muodostetut sanaesiintyvyydet.

Taulukko 1: Aineiston kahden twiitin alut.

Twiitti 1	And he did it during Friday prayers?
Twiitti 2	When it comes to mass shooters:

Taulukko 2: Sanojen esiintyvyysmäärien piirrejoukko ja piirteiden arvot kahdelle twiitille.

	and	he	did	it	during	friday	prayers	when	comes	mass	shooters
Twiitti 1	1	1	1	1	1	1	1	0	0	0	0
Twiitti 2	0	0	0	1	0	0	0	1	1	1	1

Taulukossa 2 listataan omana sarakkeenaan jokainen sana, joka esiintyy mainittujen twiittien aluissa (taulukko 1) vähintään kerran. Sana ”it” esiintyy molemmissa esimerkeissä, muut sanat puolestaan vain jommassakummassa. Sanaesiintyvyyksien käyttäminen koneoppimisalgoritmien piirteinä perustuukin yhteneväisyyksien etsimiseen sanaesiintyvyyksistä dokumenttien välillä. Kun dokumentteja on enemmän ja ne ovat pidempiä, sanaesiintyvyysmääristä voi löytyä huomattavia eroja ja samankaltaisuuksia eri dokumenttien välillä.

Piirteiden lisäksi koneoppimisen yksi olennaisimmista käsitteistä on *vaste* (*correct output* tai vain *output*) (Kotsiantis ym. 2006). Vaste voi olla joko jatkuva, kategorinen tai binäärinen. *Jatkuvalla vasteella* tarkoitetaan arvoa, joka voi saada minkä tahansa numeroarvon, kuten esimerkiksi lineaarisessa regressiossa (James ym. 2014, 59–71). *Binäärisellä vasteella* taas on vain kaksi mahdollista arvoa, esimerkiksi ”hyvä” ja ”huono”. *Kategorinen vaste* on eräänlainen jatkuvan ja binääriseen vasteen välimuoto, sillä se voi saada useita eri arvoja, joskaan ei äärettömästi. Esimerkiksi erilaiset lajiluokittelut (esim. kalalajit) sekä yleisesti kyselylomakkeissa käytettävä Likertin 5-portainen asteikko ovat koneoppimisessa kategorisia vasteita.

Piirteet ja vasteet liittyvät toisiinsa ohjaamattoman ja ohjatun koneoppimisen käsitteiden kautta. Sekä ohjatussa että ohjaamattomassa koneoppimisessa ollaan kiinnostuneita käsillä olevan mallinnusongelman muotoilusta vasteiden etsimiseksi dokumenteille. Ohjatuksi koneoppimiseksi kutsutaan tilannetta, jossa saatavilla on koneoppimismallin opetusta varten aineistoa, jossa jokaista dokumenttia vastaa tiedetty vaste. Ohjattua koneoppimista käytetään esimerkiksi usein sentimenttianalyysiin (Pang ym. 2002; Liu & Zhang 2013), jossa luokitellaan dokumentteja tai sanoja erilaisilla tunneskaaloilla. Pang, Lee ja Vaithyanatha kehittivät koneoppimismallin elokuva-arvioiden luokittelemiseen positiiviseksi, neutraaliksi tai negatiiviseksi, hyödyntäen aineistonaan IMDb -elokuvatietokannan tekstimuotoisia elokuva-arvioita sekä niissä mukana olevia numero- ja tähtiluokituksia (Pang ym. 2002). Tässä esimerkissä dokumentteja ovat siis elokuva-arviot, kategorisia vasteita sentimenttiluokitukset. Ohjatun koneoppimisen erityispiirre esimerkissä on elokuvasivuston numero- ja tähtiluokitusten käyttäminen mallin opetuksessa. Vastaavan sentimenttiluokittimen opettaminen Christchurch-aineistollemme edellyttäisi, että aineiston twiittejä luokiteltaisiin laadullisesti sentimentteihin opetusdataksi.

Ohjaamattomaksi koneoppimiseksi kutsutaan tilannetta, jossa vasteita ei ole saatavilla tai niitä ei käytetä mallin opetuksessa. Eräs esimerkki ohjaamattomasta koneoppimisesta on aihemallinnus (Blei ym. 2003), jossa piirteet lasketaan sanamääriin perustuen ja vasteita ovat aiheiden ominaisuudet. Aihemallinnuksen periaate on se, että aiheet tunnistetaan aineistosta ilman ihmisen tekemiä aiheluokituksia. Ohjaamaton koneoppiminen ja aihemallinnus sen esimerkkinä ovat suosittuja tutkimusmenetelmiä ja -kohteita, koska niitä soveltamalla vältetään työläs opetusdatan tuottaminen.

Aihemallinnuksen perusperiaatteet

Latent Dirichlet Allocation (LDA), johon tässä artikkelissa viitataan *aihemallinnuksena*, on ohjaamaton koneoppimismenetelmä (Blei ym. 2003), jota käytetään usein mediaaineistojen tutkimuksessa. Aihemalli on hierarkkinen ja generatiivinen Bayes-malli[2], jossa aineiston perusyksikköjä ovat dokumentit ja sanat. Christchurch-aineistossa dokumentit ovat twiittejä, sanat puolestaan twiittien sisältämiä sanoja.

Aihemallinnus olettaa, että aineistosta tunnistettavat aiheet selittävät jokaisen sanan esiintymistä (Puschmann ym. 2016). Aihemallinnus arvioi datasta kahta olennaista jakaumaa: aiheiden jakaumaa dokumenteissa sekä sanojen jakaumaa aiheissa. Aihemallinnuksen oletuksia ovat siis, että jokainen dokumentti on eräänlainen aiheiden yhdistelmä ja että jokaisella sanalla on tietty todennäköisyys kuulua jokaiseen aiheeseen (emt.). Käytännössä aihemallinnus käsittelee aiheita siis eräänlaisina aineiston piilevinä ominaisuuksina, joita yritetään mallinnuksessa päätellä olettaen, että samassa aiheessa merkitsevät sanat esiintyvät useammin keskenään kuin eri aiheissa merkitsevät sanat.

LDA:ssa mallin oletuksena on, että dokumentti on ikään kuin synteettisesti luotu sana kerrallaan siten, että jokaista sanaa luotaessa ensin valitaan aihe (Blei ym. 2003).

Aiheen valinnan jälkeen valitaan sana kyseisestä aiheesta, todellisuudessa siis aiheen todennäköisyysjakaumasta eri sanojen valinnalle. Myös aihetta valittaessa valinnan perusteena on aiheiden todennäköisyysjakauma. LDA:ssa on useita sisäisiä parametreja, joiden arvot määrittyvät mallinnusprosessissa datasta jakaumia estimoitaessa (emt.). Ainoa LDA:n ulkoinen, eli mallin käyttäjän sille antama parametri, on haluttu aiheiden määrä. Toisin sanoen sekä sanojen jakaumat aiheissa että aiheiden jakaumat dokumenteissa riippuvat annettujen aiheiden määrästä. Esimerkiksi, jos aiheita on hyvin paljon, joukossa saattaa olla paljon samankaltaisia, vain harvoissa dokumenteissa esiintyviä aiheita (Puschmann 2016).

Aiheiden määrän voi määritellä useilla eri tavoilla, joko laadullisesti sisältöä tarkastelemalla tai tilastollisilla menetelmillä (Nelimarkka 2019). Usein aiheiden määrä valitaan vertailemalla muutamia aiheita ja valitsemalla aihemääräksi silmämääräisesti parhaiten tulkittavat aiheet tuottava määrä (emt.). Aiheiden määrän voi valita myös aiheiden *yhtenäisyyttä* (*coherence*) mittaamalla. Tätä on tehty niin käyttäjäkokeilla kuin laskennallisestikin, yhtenäisyyden tarkoittaessa kulloinkin hieman eri asioita, esimerkiksi aiheiden tulkittavuutta tai aiheiden merkitsevimpien sanojen esiintyvyyttä toistensa kanssa dokumenteissa (Röder ym. 2015; Newman ym. 2010).

Newman, Lau, Grieser ja Baldwin (2010) ovat mitanneet aiheiden yhtenäisyyttä kokeella, jossa he pyysivät käyttäjiä arvioimaan aiheita kolmiportaisella asteikolla hyödyllisestä hyödyttömään aiheeseen. Aiheilla he tarkoittivat jokaisen aiheen kymmentä merkitsevintä sanaa, ja hyödyllisyyden he määrittivät tulkittavuutena ja helppoutena kuvailla aihetta yhdellä sanalla. He vertailivat käyttäjäkokeiden tuloksia useisiin tilastollisiin yhtenäisyysmittareihin. Tarkimmin käyttäjäkokeita vastaavaksi mittariksi osoittautui *pisteittäinen keskinäisinformaatio* (*pointwise mutual information* eli PMI). PMI on laajalti käytetty informaatioteoreettinen mittari, joka aiheiden tapauksessa mittaa sitä, kuinka usein yksittäisen aiheen yleisimmät sanat esiintyvät toistensa kanssa dokumenteissa verrattuna siihen, että ne esiintyvät dokumenteissa erikseen (Ward & Hanks 1989; Röder ym. 2015; Newman ym. 2010). Yhtenäisyyden operationalisointi PMI:n mittaamiseksi perustuu siis siihen, että yhtenäisissä aiheissa aiheiden merkitsevimmät sanat esiintyvät useammin yhdessä kuin erillisissä dokumenteissa.

Aihemallinnusta on sovellettu yhteiskuntatieteissä niin Suomessa (Ylä-Anttila 2018a; Nelimarkka 2019) kuin kansainvälisestikin (Bonilla & Grimmer 2013; Ghosh & Guha 2013; Pashakhin 2016). Esimerkiksi Bonilla ja Grimmer sovelsivat aihemallinnusta tutkiakseen Yhdysvaltain hallituksen terrorihälytysten vaikutusta siihen, mitkä aiheet saavat huomiota mediassa (Bonilla & Grimmer 2013). He yhdistivät tutkimusasetelmassaan noin 50 000 uutisartikkelin aihemallinnuksen yleisökyselyihin. He havaitsivat terrorihälytysten kasvattavan niin median terrorismiaiheen käsittelyä kuin yleisön kokemusta terrori-iskun mahdollisuudesta. Bonilla ja Grimmer kuvailevat valinneensa aiheiden määrän sekä laadullisilla että tilastollisilla perusteilla (Bonilla & Grimmer 2013). Suomessa aihemallinnusta media-aineistolle on soveltanut esimerkiksi Ylä-Anttila (2018), joka tutki maahanmuuttovastaisten MV-lehden ja Hommaforumin teksteissä esiintyvää tiedon, vastatiedon ja salaliittoteorioiden kehystystä. Hän valitsi aiheiden määrän tarkastellen kolmen eri mallin tilastollista sopivuutta aineis-

toon, mutta rajoittaakseen aiheiden kuvailuun tarvittavaa laadullista työtä hän kuitenkin päätyi pienempään aihemäärään kuin mitä tilastollinen vertailu osoitti sopivimmaksi.

Aihemallinnuksen prosessi Christchurch-aineistolle

Toteutimme aihemallinnuksen Christchurch-aineistomme englanninkielisille alkuperäistwiiteille yhteensä 15 päivän ajalta. Aihemallinnuksen prosessimme sisälsi yleisesti määritellyt vaiheet: aineiston esikäsittelyn, mallinnuksen sekä tulosten tulkinnan (Nelimarkka 2019).

Aineistomme esikäsittely sisälsi englanninkielisten twiittien tunnistamisen, näistä ylimääräisten merkkien ja hukkasanojen poiston, sekä lopulta sanojen perusmuotoon muuttamisen eli lemmauksen. Käytimme englanninkielisten twiittien tunnistamiseen Python-ohjelmointikielelle kehitettyä langdetect-kirjastoa (Danilak 2016). Kirjasto perustuu Googlen kehittämään koneoppimismalliin, joka puolestaan käyttää Wikipedian aineistoa. Yhteensä englanninkielisiä alkuperäistwiittejä tunnistettiin 148 816 kappaletta.

Hukkasanoilla tarkoitetaan luonnollisen kielen käsittelyssä sellaisia sanoja, joiden esiintyvyys tekstissä on suuri, mutta joilla on pieni merkitys analyysille (Wilbur & Sirotkin 1992). Mikäli tutkimuksen erityisenä kiinnostuksen kohteena eivät ole esimerkiksi käytetyt pronominit tai partikkelit, usein ne sisällytetään hukkasanoihin. Käytimme NLTK-ohjelmistokirjaston[3] valmista 213 englanninkielisen sanan hukkasanalistaa, joihin kuului esimerkiksi pronomineja, partikkeleita ja verbien ”be” ja ”have” eri muotoja. Hukkasanojen poistamisen lisäksi aineistosta poistettiin välimerkit ristikkomerkkiä (#) lukuun ottamatta.

Lemmauksella tarkoitetaan sanan taivutuspäätteiden poistamista ja tämän lisäksi sanan perusmuotoon muuttamista (Balakrishnan & Lloyd-Yemoh 2014). Lemmaus on erityisen hyödyllistä, kun halutaan laskea sanojen esiintyvyysmääriä, sillä sanan eri taivutusmuodot voidaan tunnistaa samaksi sanaksi. Käytimme analyysissä NLTK-kirjaston lemmausalgoritmia[4]. Lemmauksen jälkeen aineisto muutettiin Term Frequency-Inverse Document Frequency (TF-IDF) menetelmällä painotettuun muotoon, joka on tapa esittää sanojen esiintyvyydet siten, että usein koko aineistossa esiintyvät sanat kuten ”the” saavat dokumenttikohtaisesti pienemmän painoarvon kuin aineistossa yleisesti harvemmin esiintyvät sanat (Spärck Jones 1972).

Aineiston esikäsittelyn jälkeen analysoimme tilastollisesta näkökulmasta sopivaa aiheiden määrää (Röder ym. 2015). Käytimme sekä aihemallinnuksessa että aiheiden yhtenäisyyden mittauksessa Python-ohjelmointikielelle kehitettyä Gensim-kirjastoa (Řehůřek 2019). Ajoimme Röderin, Bothin ja Hinneburgin (2015) kehittämän tilastollisen yhtenäisyysmittauksen jokaiselle aiheiden määrälle välillä 1–100, ja havaitsimme tilastollisesti yhtenäisimmän aihemäärän olevan 15 aihetta. Valitsimme suurimmaksi mahdolliseksi aiheiden määräksi 100 rajataksemme työmäärää aiheiden tulkitsemisessa, mutta teoriassa tilastollisesti yhtenäisin aiheiden määrä voisi olla myös yli 100.

Aihemallinnuksen tulokset

Koska aihemallinnus olettaa jokaisen dokumentin koostuvan aiheiden yhdistelmästä, aihemallilta voidaan pyytää jokaista aihetta eniten edustavat dokumentit. Tulkitsimme aihemallinnuksen tuloksia analysoimalla jokaisen 15 aiheen 20 merkitsevintä sanaa sekä 50 edustavinta dokumenttia. Näistä muodostimme seuraavat aihekuvaukset:

Taulukko 3: Aihemallinnuksen tuloksena olevien aiheiden lyhyet kuvaukset sekä avainsanat.

	Aiheen kuvaus	Viisi merkitsevintä sanaa
1	Pakistanilaisen Naeem Rashidin marttyyrikuolema	hero, pakistan, western, inna, save, uk, indian, bbc, amp, hate
2	Numeroita sisältävät twiitit	year, 2, old, 1, 3, news, 4, young, sky, coverage
3	Kehotukset lukea muita linkkejä, twiittejä tai uutisia	read, thread, #eggboy, #hellobrother, important, mean, fuck, reason, talk, daily
4	Tunteelliset sympatian osoitukset uhreille sekä muslimiyhteisölle	christchurch, family, muslim, amp, community, love, prayer, #christchurch, victim, attack
5	Islamin korostaminen rauhaa rakastavana uskontona	#christchurch, #newzealand, #newzealandterroristattack, #newzealandshooting, #christchurchmosqueattack, religion, world, word, muslim, peace
6	Ei selkeää yhtä kuvausta, sisältää esimerkiksi pääministeri Jacinda Ardernin johtajuuden kunnioitusta	true, morning, #newzealandmosqueattacks, cannot, tweet, 🙏, address, #christchurchmosqueshooting, tomorrow, quran
7	Hyökkääjän terrorististatuksen korostaminen	he, sick, terrorist, mentally, liberal, everything, cnn, ameen, dangerous, #cdnpoli
8	Islamofobian arvostelu	white, kill, muslim, trump, christian, supremacist, terrorist, amp, realdonaldtrump, supremacy
9	Iskun välitön uutisointi ja spekulointi onko kyseessä terroriteko	new, mosque, christchurch, zealand, attack, shoot, victim, shooting, terror, massacre
10	Ei selkeää yhtä kuvausta, sisältää esimerkiksi äärioikeistoon ja salaliittoihin liittyviä twiittejä	via, youtube, too, march, #newzealandterrorattack, school, nzherald, 2019, comment, 15
11	Hyökkääjän iskusta kuvaaman videon levittämisen kieltäminen	christchurch, like, people, video, get, see, make, want, say, know
12	Uuden-Seelannin pikainen puoliautomaattisten aseiden kielto	gun, ban, law, weapon, day, change, week, use, rifle, assault
13	Kyseenalaistaminen hyökkääjän terroristiksi kutsumatta jättämisestä, kananmunan heittäminen australialaisen senaattorin päälle	terrorist, muslim, #christchurch, attack, call, medium, amp, blame, shooter, terrorism
14	Uhreille sekä rahan että uskonnollisen hyvän myöntäminen tai lahjoittaminen	allah, may, grant, victim, #prayforchristchurch, help, donate, #prayfornewzealand, family, jannah
15	Tekijän oikeudenkäynti, uhrien vertaaminen Uuden-Seelannin alkuperäisasukkaisiin	christchurch, tarrant, brenton, link, charge, control, remove, miss, father, son

Taulukossa 3 kuvatuista aiheista voidaan huomata, että ne eivät kaikissa tapauksissa ole selkeitä temaattisia kokonaisuuksia. Aiheiden merkitsevimmät sanat on suodatettu taulukkoon valitsemalla ne sanat, jotka edustavat tilastollisesti eniten kyseistä aihetta. Merkitsevimmät sanat perustuvat siis ainoastaan aihemallinnuksen päättelemiin piileviin aiheisiin. Tutkija ei näin ollen voi ennalta tietää, saako hän aihemallinnuksen tuloksena lainkaan tulkittavia aiheita.

Esimerkiksi aiheessa 13 yhdistyvät sen kyseenalaistaminen, miksi hyökkääjää ei kutsuta terroristiksi, sekä erilaiset kuvailevat twiitit tilanteesta, jossa australialaista senaattoria heitettiin kananmunalla, motiivina senaattorin esittämät muslimivastaiset kommentit. Jotkut aiheet taas liikkuvat enemmän metatasolla ja kuvaavat muiden asioiden uudelleenkehystämistä tai mediatapahtuman asettamista ajalliseen kontekstiin. Esimerkiksi suurinta osaa aiheen 3 edustavimmista twiiteistä yhdistää kehottaminen lukemaan twiittiin sisällytetyn linkin sisältö, linkkien vaihdellessa. Aiheen 9 edustavimmissa twiiteissä taas kerrotaan iskun välittömästä ajallisesta läheisyydestä erilaisilla kuvaavilla määreillä kuten ”Developing story” tai ”Breaking”.

Koska aiemman tutkimuksen perusteella sopivan aihemäärän sekä muiden mallinnuksen yksityiskohtien päättämisestä ei vallitse yksimielisyyttä (Nelimarkka 2019), tulkitsemme aihemallinnuksen tuloksia enemmänkin aineiston tiivistämisen kuin aiheiden todellisuutta kuvaavuuden näkökulmasta (Pääkkönen & Ylikoski, tulossa). Vaikka aihemallinnuksen tuottamat aiheet ja aiheiden sanalistat voivatkin vaihdella, aihemallinnuksen tuottamat aiheiden edustavimmat twiitit ovat kuitenkin aitoja otteita aineistosta. Tunnistamalla aineistosta aiheita ja tarjoamalla pääsyn näitä aiheita edustaviin twiitteihin, aihemallinnus tarjoaa lähtökohtia hybridin mediatapahtuman laadulliselle tutkimukselle. Aihemallinnuksen tuottamien aiheiden edustavimmat twiitit antoivat tietoa paitsi siitä, millainen Christchurchin iskuihin liittyvä sisältö kiertää, myös siitä miten sisältöä kierrätetään. Joissakin aiheissa nousi esiin hieman eri muodoissa kierrätettyjä lauseita, joilla osoitettiin esimerkiksi muslimiyhteisölle tai uhreille sympatiaa. Taulukossa 4 näkyy esimerkinomaisesti kaksi twiittiä, joissa ilmaistaan suuttumusta hyökkääjää ystävällisesti tervehtineen uhrin kohtalosta. Taulukon kahta twiittiä erottavat vain hashtagit ja niiden paikat, erilaiset välimerkit, hymiöt, kirjainkoot sekä eri twiitteihin liitetyt kuvat.

Taulukko 4: Kaksi aineiston twiittiä, joissa on kierrätetty samaa sisältöä mutta joista löytyy tyylieroja.

Twiitti 1	#HelloBrother He got greeted in a brotherly manner But his reply was ammunition who’s the terrorist now!!!!!!!! when islam teaches peace and love 💔💔(kuva)
Twiitti 2	He got greeted in a brotherly manner but his reply was ammunition who’s the terrorist now ?? When Islam teaches peace and love💔 #NewZealandTerroristAttack #HelloBrother (kuva)

Huomionarvoista niin näissä kuin muissakin vastaavissa esimerkeissä on se, että samankaltaiset twiitit eivät ole toistensa uudelleentwiittauksia (retweet) vaan käyttäjät ovat kopioineet tekstit toisiltaan. Näiden twiittien kierrättäminen ei siis olisi tullut yhtä selvästi näkyviin uudelleentwiitatun sisällön tarkasteluun keskittymällä. Samankaltaisten twiittien tunnistaminen on eräs aihemallin affordanssi, mutta aihemallinnus on vain yksi tapa tunnistaa samankaltaisia twiittejä.

Kiertävän sisällön lisäksi aihemallinnuksen tulokset tuottavat hybridin mediatapahtuman tutkimukselle tietoa siitä, millaisilla käytännöillä sisältöä jaetaan. Aihemallinnusta on aiemmin operationalisoitu kehystämisen tarkasteluna (Ylä-Anttila ym. 2018), ja myös aineistollamme tuloksissa ilmeni viitteitä uudelleenkehystämisestä. Taulukossa 5 on kaksi esimerkkiä aiheen 3 edustavimmista twiiteistä, joista molemmissa otetaan kantaa twiitteihin sisällytetyn linkin sisältöön. Twiitissä voidaan korostaa esimerkiksi linkin sisällön tärkeyttä (twiitti 1) tai vastenmielisyyttä (twiitti 2).

Taulukko 5: Kaksi aihetta 3 edustavaa esimerkkitwiittiä.

Twiitti 1	Important thread to read. (linkki)
Twiitti 2	This is the most disgusting statement I've ever read (linkki)

Aihemallinnuksen valintojen vaikutus mediatapahtuman tutkimukseen

Useat tekijät vaikuttavat lopputulokseen silloin, kun aihemallinnusta sovelletaan hybridin mediatapahtuman tutkimukseen. Esimerkiksi aineiston puhdistus analyysin alkuvaiheessa ei ole vain mekaaninen suorite, vaan siinä tehdyt valinnat vaikuttavat myös aihemallinnuksen lopputuloksiin. Tästä eräs esimerkki on oman analyysimme tuloksissa: aiheiden merkitsevimmissä sanoissa esiintyy sana ”amp” (esim. aihe 1), joka on puhdistettu versio merkkijonosta ”&”. Merkkijono ”&” on kuitenkin eräänlainen erikoiskoodi tarkoittamaan merkkiä ”&”, ja se esiintyy silloin tällöin twiiteissä sellaisenaan. Päätimme kuitenkin olla sisällyttämättä sitä hukkasanalistaan, koska emme voineet olla varmoja sen sekä muiden, tuntemattomien erikoiskoodien yleisyydestä twiiteissä. Olimme lisäksi päättäneet käyttää yleisesti käytössä olevaa NLTK-kirjaston hukkasanalistaa ja välimerkkien poistoa. Havaitessamme ”amp”-sanan aihemallinnuksen tuloksissa, päätimme säilyttää sen, koska emme voineet olla varmoja siitä, mitä muita Twitterille spesifejä merkkijonoja aineistosta löytyisi sattumalta tai miten voisimme varmistua kaikkien tällaisten merkkijonojen löytymisen. Tässä ratkaisussa oli siis sekä hyviä että huonoja puolia: Yhtäältä käytimme yleisesti hyväksyttyä ja toistettavaa tapaa aineiston puhdistamiseen, vaikka se toisaalta ei soveltunut täydellisesti aineistoomme tuottaen joka aiheelle koherenttia määritelmää. Käytännössä yksittäisen sanan olemassaolon merkitys aihemallinnuksessa ei ole suuri, mutta eräs ratkaisu voisi olla rakentaa hukkasanalistoja datalähdekohtaisesti

siten, että esimerkiksi Twitterissä usein esiintyvät hukkasanat olisivat laajemmassa tiedossa.

Toinen tutkimukseen vaikuttava tekijä on aihemäärä, jonka valinnan mahdollisuuksien teoriaa yllä käsittelimme. Koska sovelsimme tilastollista aiheiden yhtenäisyyden mittaria emmekä laadullista aiheiden tulkittavuuden arviointia, emme sanalistojen ja edustavimpien twiittien perusteella voineet nimetä kaikkia aiheita yhdellä sanalla kuvattaviksi aiheiksi. Aihemäärä valittiin siis käytännössä aiheiden tilastollisen sopivuuden eikä aiheiden tulkittavuuden perusteella.

Vaikka kaikki tuloksena saamamme aiheet eivät ole nykyisillä tiedoillamme tulkittavia, on kuitenkin mahdotonta tietää, olisiko kaikki aihetta edustavat twiitit lukemalla löytynyt jokin yhtenäinen narratiivi. Jos olisimme valinneet useiden tutkimusten (Ylä-Anttila 2018; Levy & Michael 2014) käyttämän tavan valita silmämääräisesti ilmeiseltä näyttävä tai muulla tavalla paras aihemäärä, valittua aihemäärää ei olisi voinut perustella yksiselitteisesti siten, että aihemäärä olisi sama esimerkiksi eri projektimme tutkijoiden valitsemana. Samalla on muistettava, että myös erilaiset tilastolliset menetelmät määrittävät aihemäärän aina joidenkin oletusten perusteella.

Aihemallinnuksen tuloksiin vaikuttava kolmas tekijä on se, miten laadulliseen luentaan otettavat dokumentit valitaan. Me valitsimme jokaisen aiheen tilastollisesti edustavimmat dokumentit, mutta olisimme voineet valita myös toisin. Esimerkiksi Ylisiurua (2017) valitsi luettavat dokumentit hyödyntämällä aiheiden avainsanoja hakusanoina siten, että luettavat dokumentit eivät välttämättä olleet aihemallissa minkään aiheen edustavimpia dokumentteja. Olisimme voineet myös valita luettavat dokumentit satunnaisotoksena esimerkiksi jokaisen aiheen tuhannesta edustavimmasta dokumentista.

Hybridin mediatapahtuman monimenetelmällinen tutkimus

Koska hybridi mediaympäristö on monimuotoinen ja -mutkainen, vaatii hybridien mediatapahtumien tutkimus monitieteellistä lähestymistapaa. Tässä artikkelissa yrityksemme on ollut hahmottaa sitä, miten monimenetelmällisesti on 1) mahdollista koota laaja aineisto 2) järjestää aineisto niin, että siitä saa mahdollisimman hyvin otetta niin laskennallisen tutkimuksen kuin digitaalisen mediaetnografiankin näkökulmasta. Tässä artikkelissamme olemme keskittyneet kuvaamaan aineiston kokoamista sekä sen järjestämistä aihemallinnuksen avulla tulkittaviksi kokonaisuuksiksi. Hanke etenee seuraavaksi tarkastelemalla erilaisia aihemallinnuksella tunnistettuja aihekokonaisuuksia ja arvioimalla näiden osuvuutta ja käyttökelpoisuutta erilaisin laadullisin menetelmin. Tulevaisuudessa häämöttää tavoite tuottaa erilaisten laskennallisten ja laadullisten menetelmien kokonaisuus, jonka avulla laajasta aineistosta voidaan muodostaa eräänlainen kartoitus (vrt. Sumiala & Harju 2019). Kartoituksen avulla olisi mahdollista tarkastella sekä niitä toimijoita, jotka keskustelussa esiintyvät, että niitä aiheita, joita he kierrättävät, sekä näiden keskinäissuhteita.

Jotta tutkijoilla olisi mahdollisuus koota laaja data-aineisto, vaatii tämä paitsi työvoimaa ja osaamista, myös infrastruktuurin, levytilaa, ohjelmistoja ja käyttöliittymiä

aineiston käsittelyyn. Havaintojemme mukaan tutkimusinfra on usein riittämätöntä media- ja viestinnäntutkimuksen alalla silloin, kun hankkeissa ei ole mukana tietotekniikan alan yksikköä, vaikka osaamista olisikin. Tämä tulisi ottaa huomioon media- ja viestinnäntutkimuksen alan tutkimusympäristöjä kehittäessä.

Kun aineisto kootaan laskennallisten ja laadullisten tutkijoiden yhteistyönä, on aineistonkeruun päättyessä olemassa jonkinlainen hahmo siitä, mitä aineisto sisältää. Tämä käsitys on kuitenkin hatara ja tarvitaan menetelmiä, joiden avulla voidaan uuttaa datasta laajempi kuva. Nämä tarjoaisivat askelmerkkejä, joita seuraten voitaisiin käydä laadullisin menetelmin syvemmälle tiettyihin aineiston ulottuvuuksiin tai tapauksiin.

Aihemallinnus on yksi mahdollinen tapa jäsentää laajaa hybridin mediatapahtuman yhteydessä koottua datamassaa. Aikaisemmissa tutkimuksissa laskennallisesti kerättyä datamassaa on jäsennetty esimerkiksi laskennallisella sisällönanalyysillä ja verkostoanalyysillä (Sumiala ym. 2016). Tässä artikkelissa olemme kuvanneet ne tekniset periaatteet, jotka vaikuttavat datankeräyksen ja koneoppimisen taustalla. Nämä lähtökohdat vaikuttavat tuloksiin yhtä lailla kuin laadullisessa tutkimuksessa tehtävät valinnat ja kontekstit.

Aihemallinnusvaiheen jälkeen luontevia askeleita sisällön kierron jatkotutkimuksessa olisi joko ottaa tarkempaan tarkasteluun aihemallinnuksessa löytyneitä kiertäviä lauseita tai selvittää koko aineistossa esiintyvät samankaltaiset twiitit. Aihemallinnus ei ole ainoa mahdollinen menetelmä samankaltaisten twiittien löytämiseksi, eräitä perinteisempiä tapoja analysoida samankaltaisia merkkijonoja ovat merkkijonojen etäisyysalgoritmit (Cohen ym. 2003).

Aihemallinnuksen tulokset tarjoavat mahdollisuuksia muodostaa sekä empiirisiä että teoreettisia tutkimuskysymyksiä jatkotutkimuksen pohjaksi. Christchurchin iskujen aineistosta kiinnostaviksi teemoiksi nousivat esimerkiksi Uuden-Seelannin pääministerin rooli sekä kysymys salaliittoteorioiden kierrosta, samoin kuin väkivaltaisen äärioikeistolaisuuden nousu sosiaalisessa mediassa. Näihin paneudumme jatkossa tarkastellen aineistoa sekä laskennallisilla että laadullisilla menetelmillä. Liitymme tässä laajaan joukkoon mediatutkijoita ja yhteiskuntatieteilijöitä, jotka tavoittelevat otetta laajoista digitaalisista media-aineistoista.

Viitteet

1 Twitter-datan keräämiseen liittyvistä eettisistä ja tekijänoikeudellisista pelisäännöistä käydään aktiivista keskustelua tätä artikkelia kirjoitettaessa. Laaksonen ja Salonen pohtivat teemaa Rajapinnassa: https://rajapinta.co/2018/12/04/kuka-saa-paattaa-mita-dataa-tutkijalla-on-kaytossaan-ei-ainakaan-amerikkalainen-suuryritys/

2 Bayes-mallit ovat tilastollisia malleja, jotka ottavat huomioon, että havainnot eivät sellaisenaan kuvaa täysin todellisuutta. Sen vuoksi ne ovatkin suosittuja epävarmuutta sisältävien prosessien mallinnuksessa (esim. Sharif-Razavian & Zollmann 2008).

3 https://www.nltk.org/

4 https://www.nltk.org/_modules/nltk/stem/wordnet.html

Kirjallisuus

Balakrishnan, Vimala & Lloyd-Yemoh, Ethel (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. Lecture Notes on Software Engineering 2:3, 262–267. Saatavilla: http://lnse.org/papers/134-I3007.pdf (luettu 1.8.2019).

Christchurch shootings: 49 dead in New Zealand mosque attacks. (2019, 15. maaliskuuta). *BBC*, Uutissivusto. Saatavilla: https://www.bbc.com/news/world-asia-47578798 (luettu 21.5.2019).

Blei, David; Ng, Andrew & Jordan, Michael (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*. 3 (4–5), 993–1022.

Bonilla, Tabitha & Grimmer, Justin (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. *Poetics 41*:6, 650–669. https://doi.org/10.1016/j.poetic.2013.06.003

Boyd, danah & Crawford, Kate (2012). Critical Questions for Big Data. *Information, Communication & Society* 15:5, 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bruno, Nicola (2011). Tweet first, verify later? How real-time information is changing the coverage of worldwide crisis events. Reuters Institute for the Study of Journalism. Oxford: University of Oxford. Saatavilla: https://nicolabruno.files.wordpress.com/2011/05/tweet_first_verify_later2.pdf (luettu 1.8.2019).

Chadwick, Andrew (2013). The Hybrid Media System: Politics and Power. Oxford University Press.

Church, Kenneth Ward & Hanks, Patrik (1989). Word association norms, mutual information, and lexicography. *Proceedings of the 27th annual meeting on Association for Computational Linguistics (ACL '89)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 76-83. https://doi.org/10.3115/981623.981633

Cioffi-Revilla, Claudio (2010). Computational social science. Wiley Interdisciplinary Reviews: Computational Statistics 2:3, 259–271. http://dx.doi.org/10.2139/ssrn.1708051

Turkish citizen hurt in Christchurch attacks dies, NZ death toll at 51: Minister. (2019, 2. toukokuuta). *CNA*, Uutissivusto. Saatavilla: https://www.channelnewsasia.com/news/world/christchurch-terror-attack-new-zealand-death-toll-new-11497782 (luettu 21.5.2019).

Cohen, William; Ravikumar, Pradeep & Fienberg, Stephen (2003). A comparison of string distance metrics for name-matching tasks. Proceedings of the 2003 International Conference on Information Integration on the Web (IIWEB'03). AAAI Press, 73–78. Saatavilla: http://dc-pubs.dbs.uni-leipzig.de/files/Cohen2003Acomparisonofstringdistance.pdf (luettu 10.2.2020).

Couldry, Nick; Hepp, Andreas & Krotz, Friedrich (toim.) (2010). *Media Events in a Global Age*. Abingdon: Routledge.

Danilak, Michal (2016). Langdetect. https://pypi.org/project/langdetect/

Dayan, Daniel. (2010). Beyond media events. Disenchantment, derailment, disruption. Teoksessa: Couldry, Nick; Hepp, Andreas & Krotz, Friedrich (toim.). *Media Events in a Global Age.*, Abingdon: Routledge, 23–42

Dayan, Daniel & Katz, Elihu (1992). *Media Events. The Live Broadcasting of History*. Cambridge, MA: Harvard University Press.

Eide, Elisabeth; Kunelius, Risto & Phillips, Angela (toim.) (2008). *Transnational Media Events: The Mohammed Cartoons and the Imagined Clash of Civilizations.* Göteborg, Sweden: Nordicom. Saatavilla: https://www.nordicom.gu.se/sites/default/files/publikationer-hela-pdf/transnational_media_events.pdf (luettu 1.8.2019).

Fiesler, Casey & Proferes, Nicholas (2018). "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 1–14. https://doi.org/10.1177/2056305118763366

Ghosh, Debarchana & Guha, Rajarshi (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science* 40:2, 90–102. https://doi.org/10.1080/15230406.2013.776210

Huhtamäki, Jukka; Russell, Martha G.; Rubens, Neil & Still, Kaisa (2015). Ostinato: The exploration-automation cycle of user-centric, process-automated data-driven visual network analytics. Teoksessa: Matei, Sorin Adam; Russell, Martha G.; & Bertino, Elisa (toim.). *Transparency in Social Media: Tools, Methods and Algorithms for Mediating Online Interactions*. Springer International Publishing Switzerland, 197–222. https://doi.org/10.1007/978-3-319-18552-1_11

James, Gareth; Witten, Daniela; Hastie, Trevor & Tibshirani, Robert (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company Incorporated.

Katz, Elihu & Liebes, Tamara (2007). 'No more peace!' How disaster, terror and war have upstaged media events. *International Journal of Communication* 1, 157–66. Saatavilla: https://ijoc.org/index.php/ijoc/article/view/44 (luettu 1.8.2019).

Kotsiantis, Sotiris; Zaharakis, Ioannis & Pintelas, Panagiotis (2006). Supervised machine learning: A review of classification techniques. *Artificial Intelligence Review* 26, 159–190. https://doi.org/10.1007/s10462-007-9052-3

Kraidy, Marwan (2005). *Hybridity. The Cultural Logic of Globalization*. Philadelphia: Temple University Press.

Kyriakidou, Maria (2008). Rethinking media events in the context of a global public sphere: Exploring the audience of global disaster in Greece. *Communications. The Journal of European Communication Research* 33:3, 273–291. https://doi.org/10.1515/COMM.2008.018

Laaksonen, Salla-Maaria; Nelimarkka, Matti; Tuokko, Mari; Marttila, Mari; Kekkonen, Arto & Villi, Mikko (2017). Working the fields of big data: Using big-data-augmented online ethnography to study candidate–candidate interaction at election time. *Journal of Information Technology & Politics* 14:2, 110–131. https://doi.org/10.1080/19331681.2016.1266981

Latour, Bruno (1993). *We Have Never Been Modern*. Cambridge, MA: Harvard University Press.

Lazer, David; Pentland, Alex; Adamic, Lada; Aral, Sinan; Barabási, Albert-László; Brewer, Devon; Christakis, Nicholas; Contractor, Noshir; Fowler, James; Gutmann, Myron; Jebara, Tony; King, Gary; Macy, Michael; Roy, Deb & Van Alstyne, Marshall (2009). Computational social science. *Science* 323:5915, 721–723. https://doi.org/10.1126/science.1167742

Levy, Karen E. C. & Franklin, Michael (2014). Driving regulation: Using topic models to examine political contention in the U.S. trucking industry. *Social Science Computer Review* 32:2, 182–194. https://doi.org/10.1177/0894439313506847

Liebes, Tamar (1998). Television's disaster marathons. A danger for democratic processes? Teoksessa: Liebes, Tamara & Curran, James (toim.). *Media, Ritual and Identity*. London: Routledge, 71–84.

Liu, Bing & Zhang, Lie (2013). A survey of opinion mining and sentiment analysis. Teoksessa: Aggarwal Charu C. & Zhai, ChengXiang (toim.). Mining Text Data. Springer, Boston, MA, 415–463. https://doi.org/10.1007/978-1-4614-3223-4_13

Lovins, Julie Beth (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11:1–2, 22–31. Saatavilla: https://pdfs.semanticscholar.org/6b38/53f08c482fe1bfbe39d656d50a8c73976f3c.pdf (luettu 1.8.2019).

Nelimarkka, Matti (2019). Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja. *Politiikka* 61.1, 6–33.

Newman, David; Lau, Jey Han; Grieser, Karl & Baldwin, Timothy (2010). Automatic evaluation of topic coherence. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Saatavilla: https://dl.acm.org/citation.cfm?id=1858011 (luettu 1.8.2019).

Nossek, Hillel (2008). 'News media'—Media events: Terrorist acts as media events. *Communications. The Journal of European Communication Research* 33:3, 313–330. https://doi.org/10.1515/COMM.2008.020

Pang, Bo; Lee, Lillian & Vaithyanatha, Shivakumar (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of EMNLP*, 79–86..https://doi.org/10.3115/1118693.1118704

Pashakhin, Sergei (2016). Topic Modeling for Frame Analysis of News Media. Artificial Intelligence and Natural Language AINL FRUCT 2016. Saatavilla: https://linis.hse.ru/data/2017/08/25/1174061267/Pas.pdf (luettu 1.8.2019).

Procter, Rob; Vis, Farida, & Voss, Alex (2013). Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology* 16:3, 197–214. https://doi.org/10.1080/13645579.2013.774172

Puschmann, Cornelius & Scheffler, Tatjana (2016). Topic Modeling for Media and Communication Research: A short primer HIIG Discussion Paper Series No. 2016-05. http://dx.doi.org/10.2139/ssrn.2836478

Pääkkönen, Juho & Ylikoski, Petri (tulossa). Humanistic interpretation and machine learning.

Řehůřek, Radim (2019): Gensim, https://radimrehurek.com/gensim/

Rothenbuhler, Eric W. (2010). Media events in the age of terrorism and Internet. *The Romanian Review of Journalism and Communication* IV:2, 34–41.

Röder, Michael; Both, Andreas & Hinneburg, Alexander (2015). Exploring the Space of Topic Coherence measures. Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6. https://doi.org/10.1145/2684822.2685324

Shalev-Shwartz, Shai & Ben-David, Shai (2014). *Understanding Machine Learning: from Theory to Algorithms*. New York: Cambridge University Press.

Sharif-Razavian, Narges & Zollmann, Andreas (2008). An Overview of Nonparametric Bayesian Models and Applications to Natural Language Processing. Technical Report. Saatavilla: http://www.cs.cmu.edu/~zollmann/publications/nonparametric.pdf (luettu 1.8.2019).

Sonnevend, Julia (2016). *Stories without Borders. The Berlin Wall and the Making of a Global Iconic Event*. New York: Oxford University Press.

Spärck Jones, Karen (1972). A statistical interpretation of term specifity and its application in retrieval. *Journal of Documentation* 28:1, 11–21. https://doi.org/10.1108/eb026526

Sumiala, Johanna & Tikka, Minttu (2020, tulossa). Digital media ethnography. A proposal for research on the move. *Journal of Digital Social Research*.

Sumiala, Johanna & Harju, Anu (2019). "No more apologies": Violence as a trigger for public controversy over Islam in the digital public sphere. *Journal of Religion, Media and Digital Culture 8*:1, 132–152. https://doi.org/10.1163/21659214-00801007

Sumiala, Johanna; Tikka, Minttu; Huhtamäki, Jukka & Valaskivi, Katja (2016). #JeSuisCharlie. Towards a multi-method study of hybrid media event. *Media and Communication* 4:4, 97–108. http://dx.doi.org/10.17645/mac.v4i4.593

Sumiala, Johanna; Valaskivi, Katja; Tikka, Minttu; Huhtamäki, Jukka (2018): Hybrid Media Events: The Charlie Hebdo Attacks and the Global Circulation of Terrorist Violence. Bingley, UK: Emerald.

Tufekci, Zeynep (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. Proceedings of the International AAAI conference on weblogs and social media. Saatavilla: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8062/8151 (luettu 1.8.2019).

Vaccari, Christian; Chadwick, Andrew & O'Loughlin, Ben (2015). Dual screening the political: Media events, social media, and citizen engagement. *Journal of Communication* 65:6, 1041–1061. https://doi.org/10.1111/jcom.12187

Valaskivi, Katja & Sumiala, Johanna (2014). Circulating social imaginaries: Theoretical and methodological reflections. *European Journal of Cultural Studies* 17:3, 229–243. https://doi.org/10.1177/1367549413508741

Valaskivi, Katja; Rantasila, Anna; Tanaka, Mikihito & Kunelius, Risto (2019). *Traces of Fukushima. Global Events, Networked Media and Circulating Emotions*. Lontoo: Palgrave Macmillan.

Vis, Farida (2013). Twitter as a reporting tool for breaking news. *Digital Journalism* 1:1, 27–47. https://doi.org/10.1080/21670811.2012.741316

Wahlquist, Calla (2019, 19. maaliskuuta). Ardern says she will never speak name of Christchurch suspect. *Guardian*, Uutissivusto. Saatavilla: https://www.theguardian.com/world/2019/mar/19/new-zealand-shooting-ardern-says-she-will-never-speak-suspects-name (luettu 21.5.2019).

Wilbur, W. John & Sirotkin, Karl (1992). The automatic identification of stop words. *Journal of Information Science* 18:1, 45–55. https://doi.org/10.1177/016555159201800106

Ylisiurua, Marjoriikka (2017). Aihemallinnuksen mahdollisuudet sosiaalisen median aineistojen jäsentämisessä – terveyskeskustelu Suomi24-verkkopalstalla. *Kulutustutkimus.Nyt* 11:2.

Ylä-Anttila, Tuukka (2018). Populist knowledge: 'Post-truth' repertoires of contesting epistemic authorities. *European Journal of Cultural and Political Sociology* 5:4, 356-388. https://doi.org/10.1080/23254823.2017.1414620

Ylä-Anttila, Tuukka; Eranti, Veikko & Kukkonen, Anna (2018). Topic modeling as a method for frame analysis: Data mining the climate change debate in India and the USA. SocArXiv. March 20. https://doi.org/10.31235/osf.io/dgc38

Zhang, Yin; Jin, Rong & Zhou, Zhi-Hua (2010). Understanding bag-of-words model: A statistical framework. *International Journal of Machine Learning and Cybernetics* 1, 43–52. https://doi.org/10.1007/s13042-010-0001-0